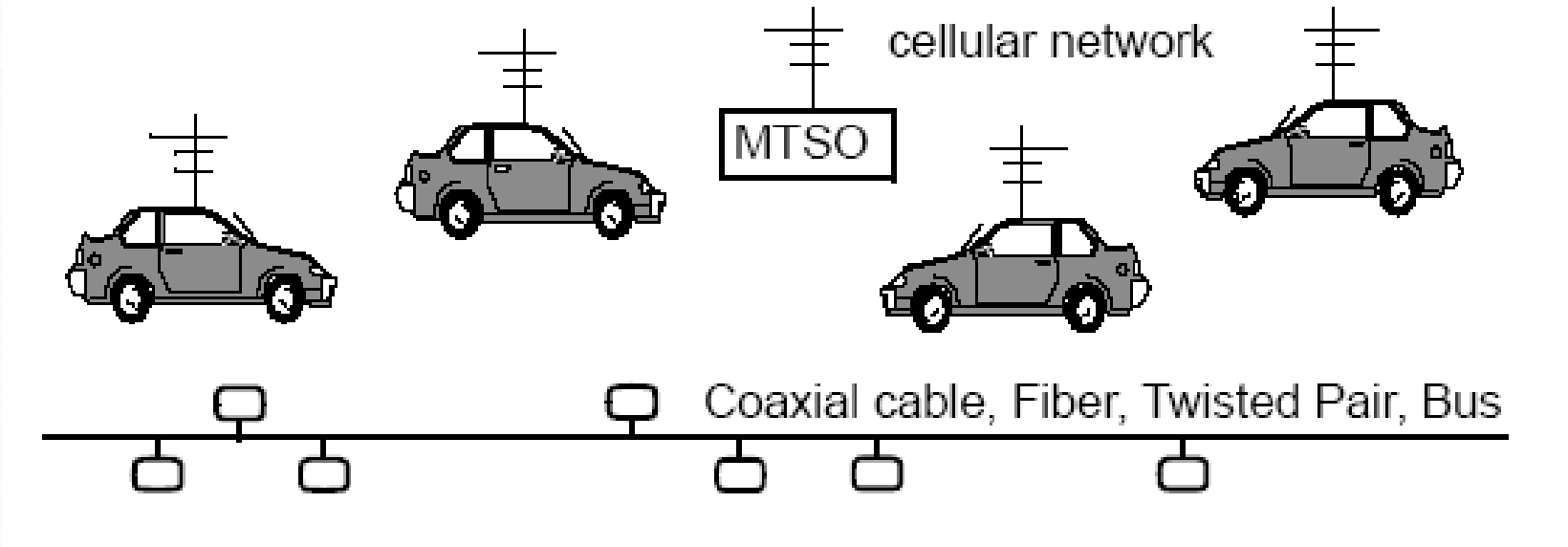
Queueing Theory and its Applications

From: Mrs. Deepali
Assistant Professor
Department of Mathematics

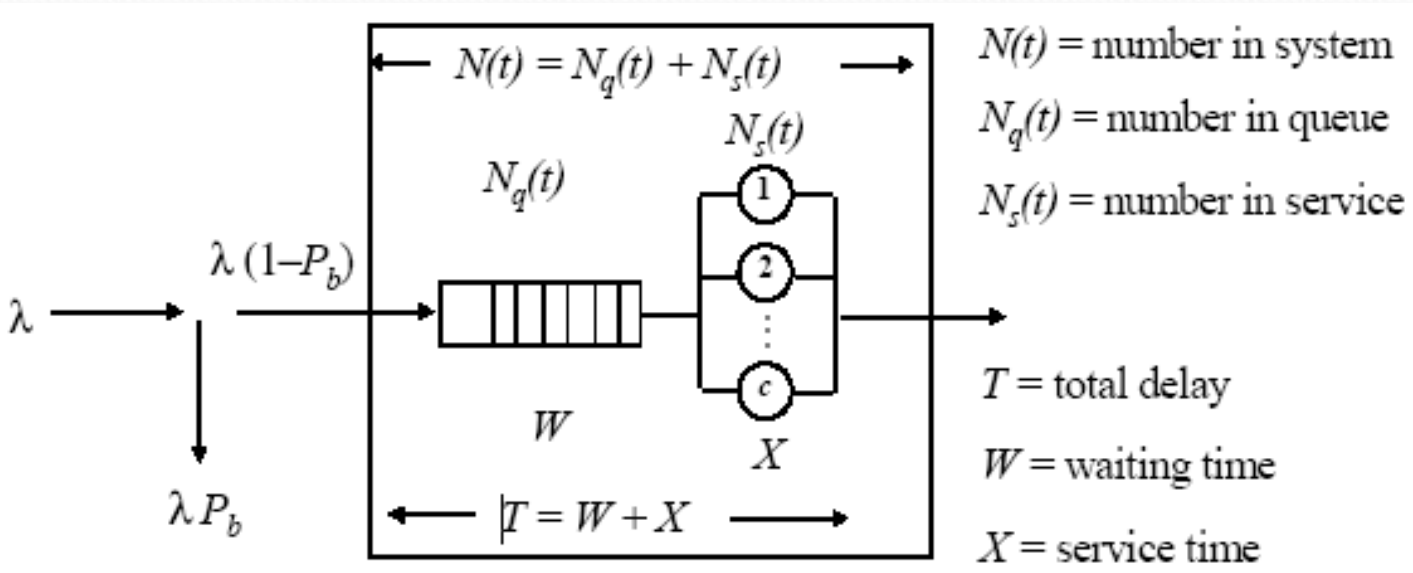
Multiaccess Problem

- How to let distributed users (efficiently) share a single broadcast channel?
  ⇒ How to form a queue for distributed users?
- The protocols we used to solve this multiaccess problem are called multiaccess protocols. They are the lower sublayer of Data Link Control layer in the OSI model.
- The queueing theory studies properties of waiting queues. The mathematical formula of queueing theory can be used to evaluate the efficiency of different queueing system designs. In our applications, the efficiency of various multiaccess protocols.



Queueing Theory

Parameter of interest to queueing analysis:
λ average (avg, or mean) arrival rate (requests/sec) (packets/sec)
μ avg service rate (requests/sec) (packets/sec)
ρ=λ/μ utilization or traffic density, the ratio of system load to system capacity
N avg no.of requests in the system, including those in buffers and in servers.
Tw avg waiting time for a request at the buffer
Ts avg service time for a request
T avg delay in the system = Ts +Tw; its inverse is the avg system throughput.
P_B probability of request lost (due to buffer full situation)



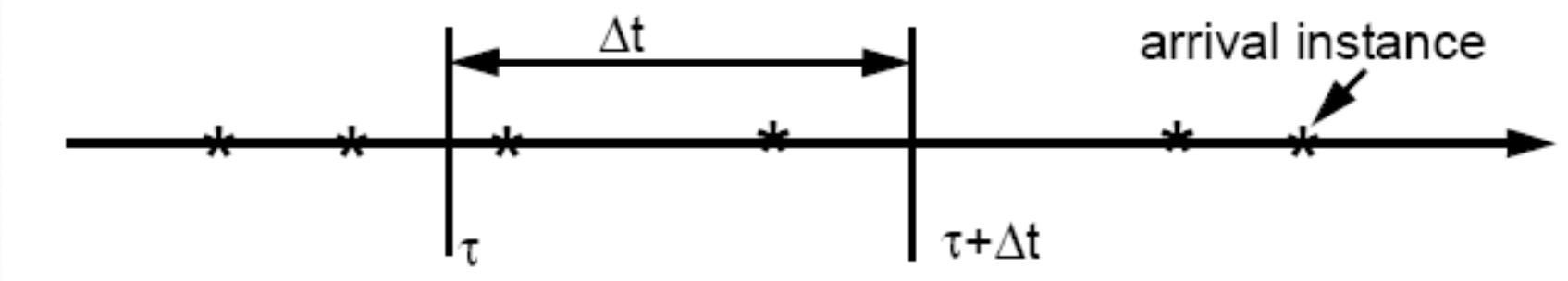
Arrival Process / Service Time / Servers / Max Occupancy

Interarrival times τ	Service times X	1 server	K customers
M = exponential	M = exponential	c servers	unspecified if
D = deterministic	D = deterministic	infinite	unlimited
G = general	G = general		
Arrival Rate:	Service Rate:		
$\lambda = 1/E[\tau]$	$\mu = 1/E[X]$		

Multiplexer Models:	M/M/1/K, M/M/1, M/G/1, M/D/1
Trunking Models:	M/M/c/c, M/G/c/c
User Activity:	M/M/∞, M/G/ ∞

Poisson Process

The random process used most frequently to model the arrival pattern.



The statistics of the Poisson process can be observed at any starting time:
1. Prob[instance occurrence in $(\tau, \tau+\Delta t)$] = $\lambda\Delta t+o(\Delta t)$; λ is mean arrival rate;
2. Prob[no instance occurrence in $(\tau, \tau+\Delta t)$] = $1-\lambda\Delta t+o(\Delta t)$;
3. Arrivals are memoryless An arrival in one time interval of length Δt is independent of arrivals in previous or future intervals.

$o(\Delta t)$ implies that other terms are higher order in Δt and approaches 0 faster than Δt. Prob[2 or more arrivals in $(\tau, \tau+\Delta t)$] = $o(\Delta t)$.

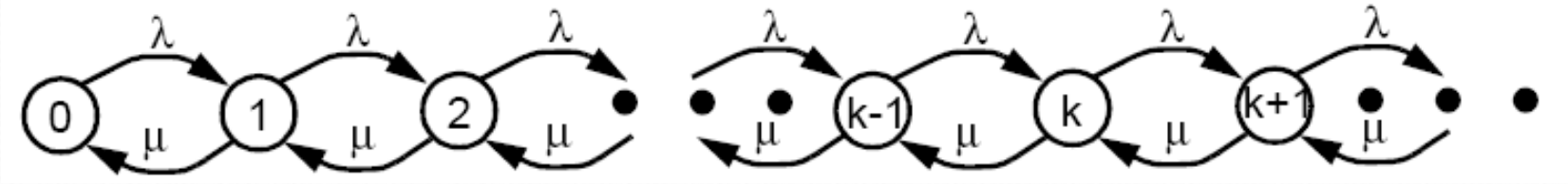
Note that the probability is independent of t.

M/M/1 Queue

Poisson arrival process, exponential service time, single server

p_k=Prob[k customers in the system]

To analyze M/M/1 Queue, let us examine its system behavior described by the following state transition diagram. State number represent the number of customers in the system



At equilibrium state the following equations hold

$$\lambda p_0 = \mu p_1 \qquad (\lambda + \mu) p_k = \lambda p_{k-1} + \mu p_{k+1} \quad k \geq 1$$

Alternatively, $\lambda p_k = \mu p_{k+1} \quad k \geq 1$

Solving $p_k = \frac{\lambda}{\mu} p_{k-1} = \rho p_{k-1}$ where $\rho = \frac{\lambda}{\mu}$ by definition $\sum_{k=0}^{\infty} p_k = 1$

We get $p_k = (1-\rho)\rho^k \qquad k \geq 0$

Mean number of customers in the system $N = \sum_{k=0}^{\infty} k p_k = \frac{\rho}{1-\rho}$

M/M/1 Queue & Little's Law

- Little's Law $N = \lambda T$ where T is the mean delay inside the system.
- Mean delay for M/M/1 system

$$T = \frac{N}{\lambda} = \frac{\rho / \lambda}{1-\rho} = \frac{1/\mu}{1-\rho} = \frac{1}{\mu - \lambda} = \frac{1}{\mu' C - \lambda}$$

where C= server service rate in # operations/sec (#bits/sec for transmission system)

= mean # operations/customer (#bits/packet for TX system)

$1/\mu'$

Evolution of Queues

multiple queues → single queue multiple servers → single queue shared server

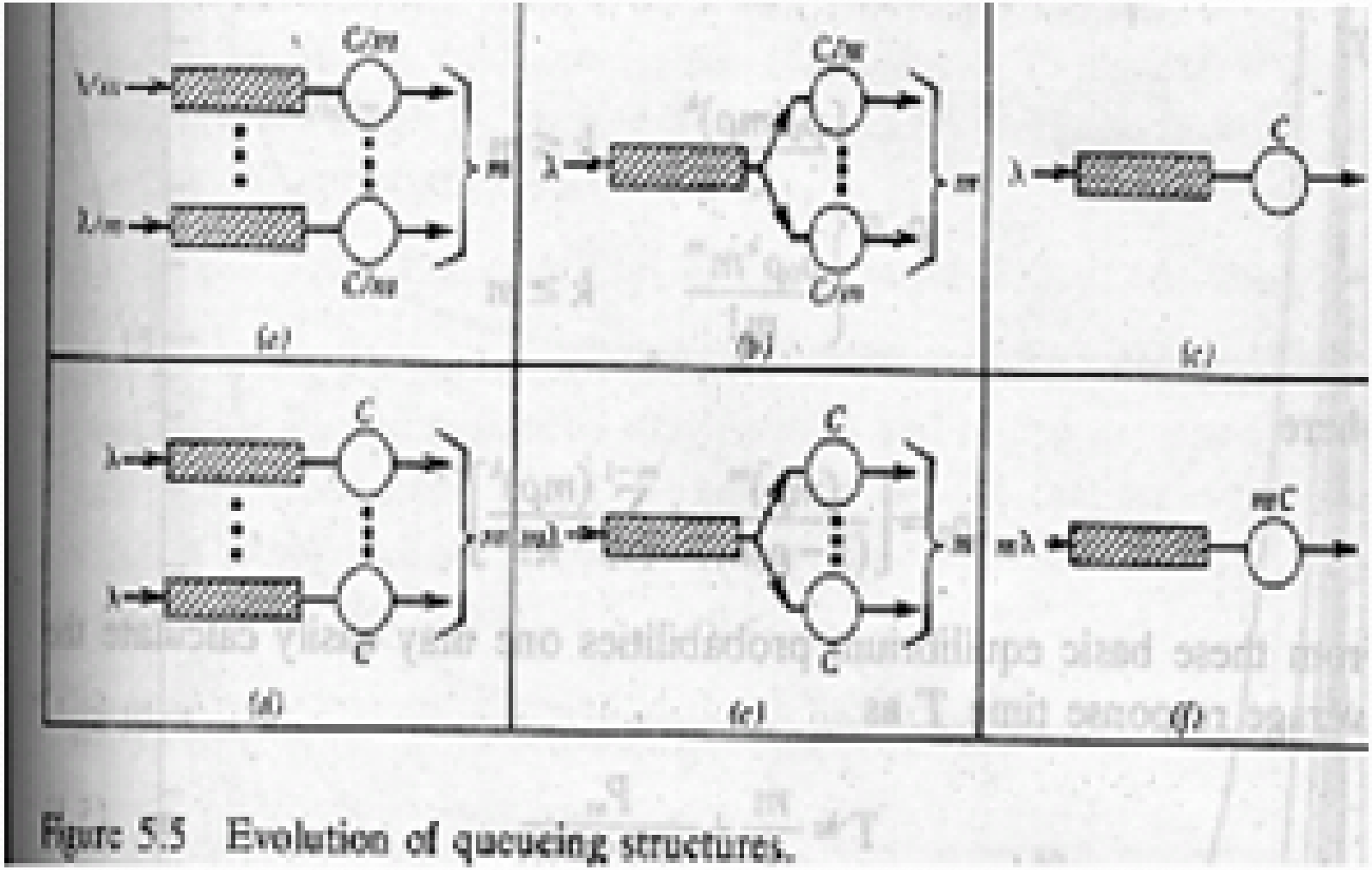


Figure 5.5 Evolution of queueing structures.

Why We Use Statistical Multiplexing?

Advanced Queueing analysis presents the following interesting result:

$$T(1, \lambda, C) \leq T(m, \lambda, C)$$

while

$$W(1, \lambda, C) \geq W(m, \lambda, C)$$

where (m,λ,C) is a system with m servers, total capacity C, arrival rat λ,
T is the total system delay and W i the waiting time in the queue.

Sharing a single high speed server increase "contention" delay in the queue but decrease overall delay due to much shorter service time.

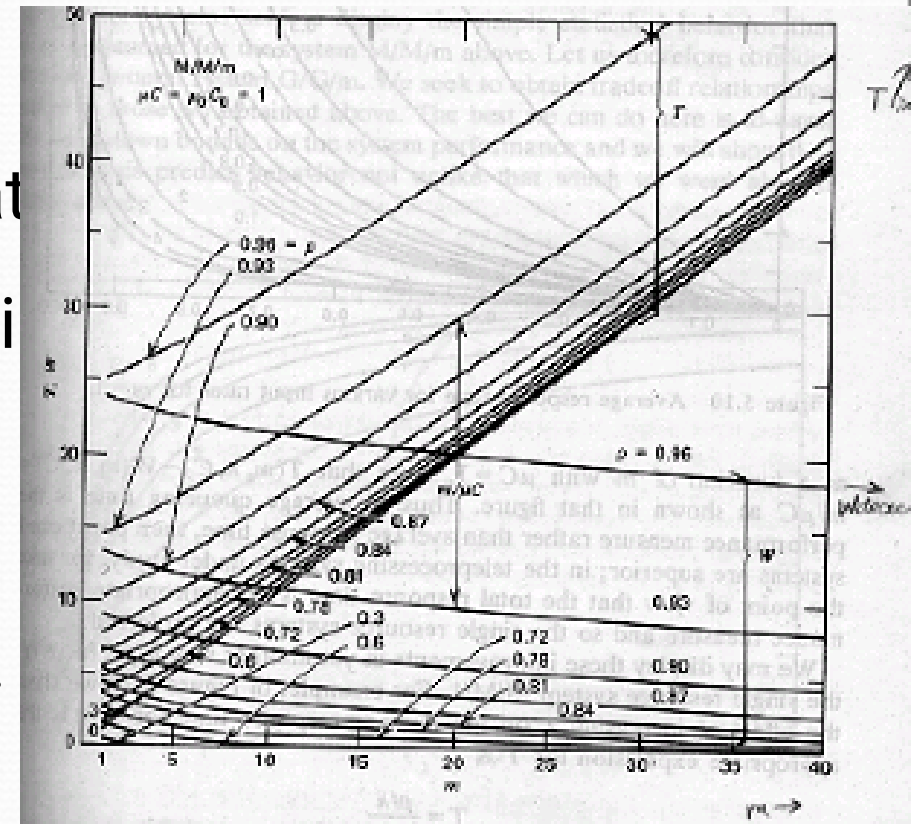
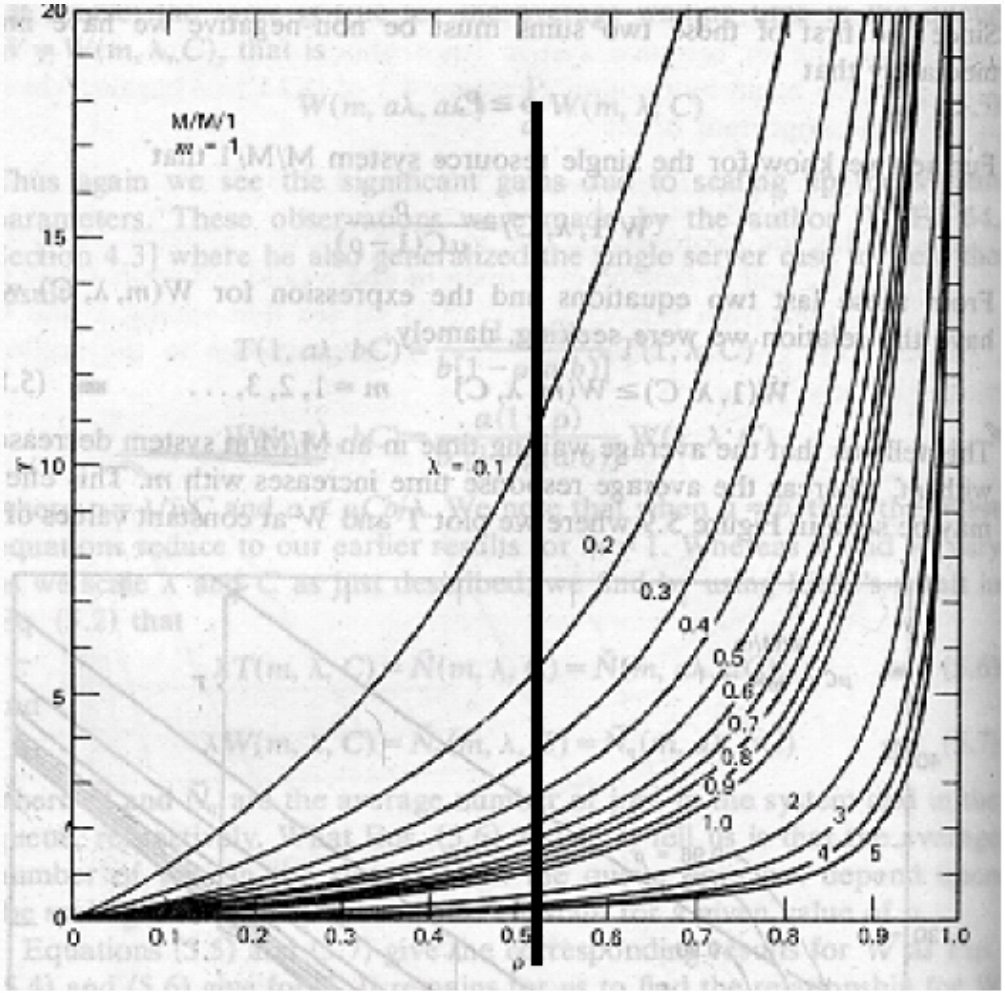


Figure 3.9 Average response time and average wait at constant loads.

Scale factor of M/M/1 Queueing System

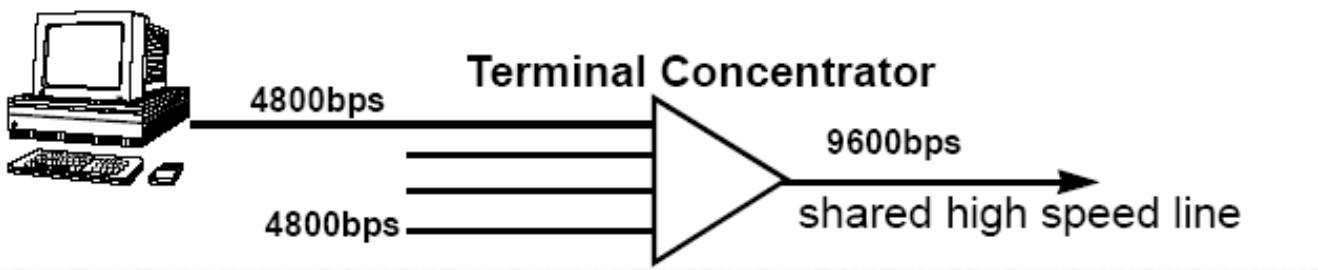
$T = \frac{\rho / \lambda}{1 - \rho}$ if λ ↑ and C ↑ so that remains constant, then T↓.

This benefit is in additional to economy of scale.



Application of Queueing Theory

Case 1. Terminal Concentrators:



mean packet length 1/μ'=1000bits/packet

input lines traffic are Poisson process with mean arrival rate λi=2 packets/sec.

Q1: What is the mean delay experienced by a packet from the time the last bit arrives at the concentrator until the moment that bit is retransmitted on the output line? $T = \frac{1}{\mu' C - \lambda}$

Use , where λ = 4xλi=8, μ'C=9.6packets/sec.

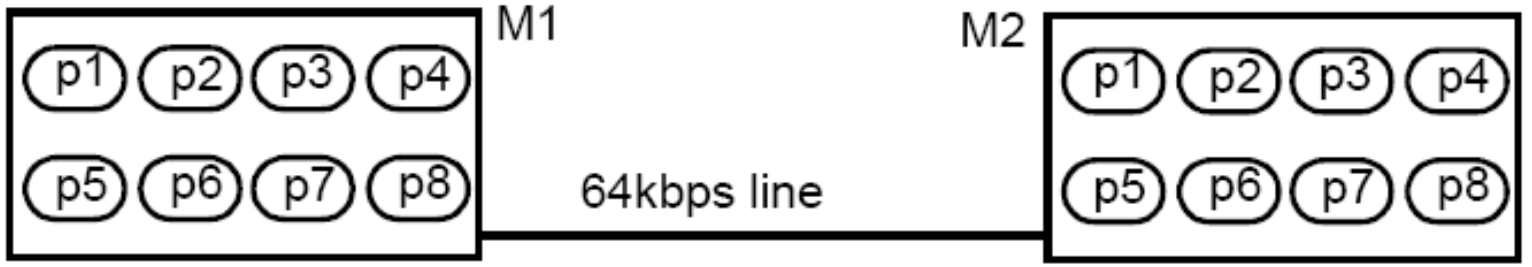
→ T=1/(9.6−8)=0.625 sec.

Q2: What is the mean number of packets in the concentrator, including the one in service?

A: Use the little's law N=λT=8x0.625=5!

Application of Queueing Theory

Dedicated vs. Shared Channels:



Eight parallel sessions using this 64kbps line. Each session generates Poisson traffic with λi=2 packets/sec. Packet lengths are exponentially distributed with a mean of 2000 bits.

There are two design choices:

a) Each session is given a dedicated 8kbps channel (via FDM or TDM).

b) Packets of all sessions compete for a single 64kbp shared channel.

Which one gives a better response time?

A: a) For 8kbps channel, λ=2packets/sec, μ'=1/2000 packets/bit, C=8000bits/sec, μ'C=4 packets/sec, T=1/(μ'C-λ)=1/(4−2)=0.5 sec.

b) For 64kbps shared channel, λ=8x2=16packets/sec, μ'=1/2000 packets/bit, C=64000bits/sec, μ'C=32 packets/sec, T=1/(μ'C-λ)=1/(32−16)=0.0667sec.

Reason?